

# Une approche hybride de l'IA pour les technologies éducatives: augmenter les STI avec l'IA générative

Sofiya Kobylanskaya<sup>1</sup>, Catherine de Vulpillières<sup>1</sup>, Pierre-Yves Oudeyer<sup>1,2</sup>

(1) EvidenceB, Paris (2) Inria, Bordeaux

sofiya-k@evidenceb.com catherine-d@evidenceb.com pierre-yves.oudeyer@inria.fr



EvidenceB



## Introduction

- Déclin scolaire y compris en math (PISA, TIMSS)
- Résultats prometteurs des STIs (Clément et al. 2015), mais grand nombre d'exercices produits manuellement
- Essor des LLMs, mais problèmes éthiques et de performance pour l'éducation, manque de données libres (Jurenka et al. 2024, Tack & Piech 2022), coût énergétique

### Solution possible

Nous proposons une approche hybride de l'IA au service de l'éducation, en combinant la personnalisation offerte par les Systèmes de Tutorat Intelligents (STI) avec de l'IA générative permettant de générer un grand nombre de contenus éducatifs de qualité, tout en respectant les contraintes pédagogiques et cognitives.

## STI d'EvidenceB

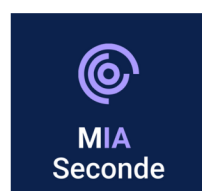
- Banque d'exercices conçus **manuellement** par des experts en pédagogie et en sciences cognitives (~ 40 000 exercices)
- Les exercices sont organisés en **graphes**
- Les exercices sont orchestrés par un **algorithme d'apprentissage par renforcement ZPDES** pour la personnalisation du parcours
- Objectif**: maximiser le progrès des élèves (Gottlieb & Oudeyer, 2018)
- Impact prouvé avec **RCT** (Roell et al. 2025a, 2025b)
- Regroupement** des élèves en clusters en fonction de leur parcours (affiché sur le tableau de bord de l'enseignant)



10 000  
exercices



10 000  
exercices



20 000  
exercices

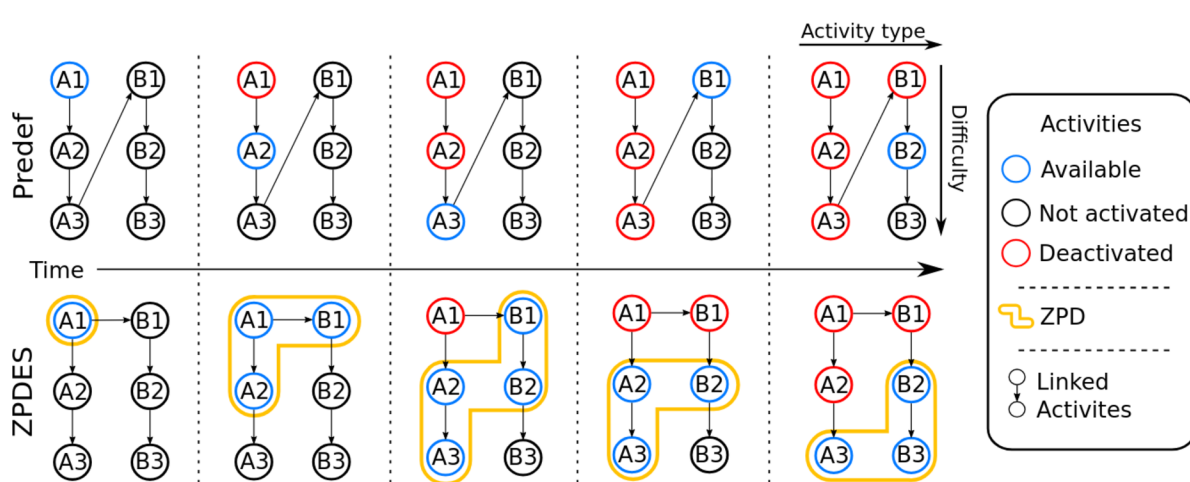
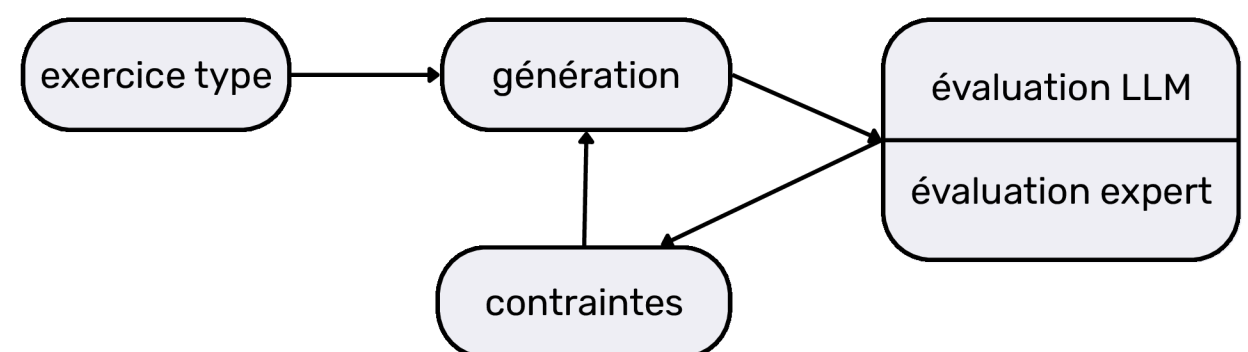


Figure 1: The space of available activities always contain only one activity in the predefined linear sequence (Predef) while the space expands over time with ZPDES to allow a diversity of exploration and find the best activities for the learner. The type of activity is represented by a letter (A or B), and the difficulty by a number (1, 2 or 3).

## Méthodologie

- Phase 1: génération du contenu
  - Conception d'exercices type par des experts (Manuels libres d'Ile-de-France ~150 exercices)
  - Reproduction des exercices types par un LLM
- Phase 2: évaluation du contenu généré
  - LLM-as-judge
  - Validation par des experts
  - Comparaison évaluation experte vs LLM-as-judge with Alt-score (Calderon et al. 2025)
- Phase 3: étude d'impact
  - RCT (Roell et al. 2025a, 2025b)

## Processus de génération



## Critères d'évaluation des exercices générés

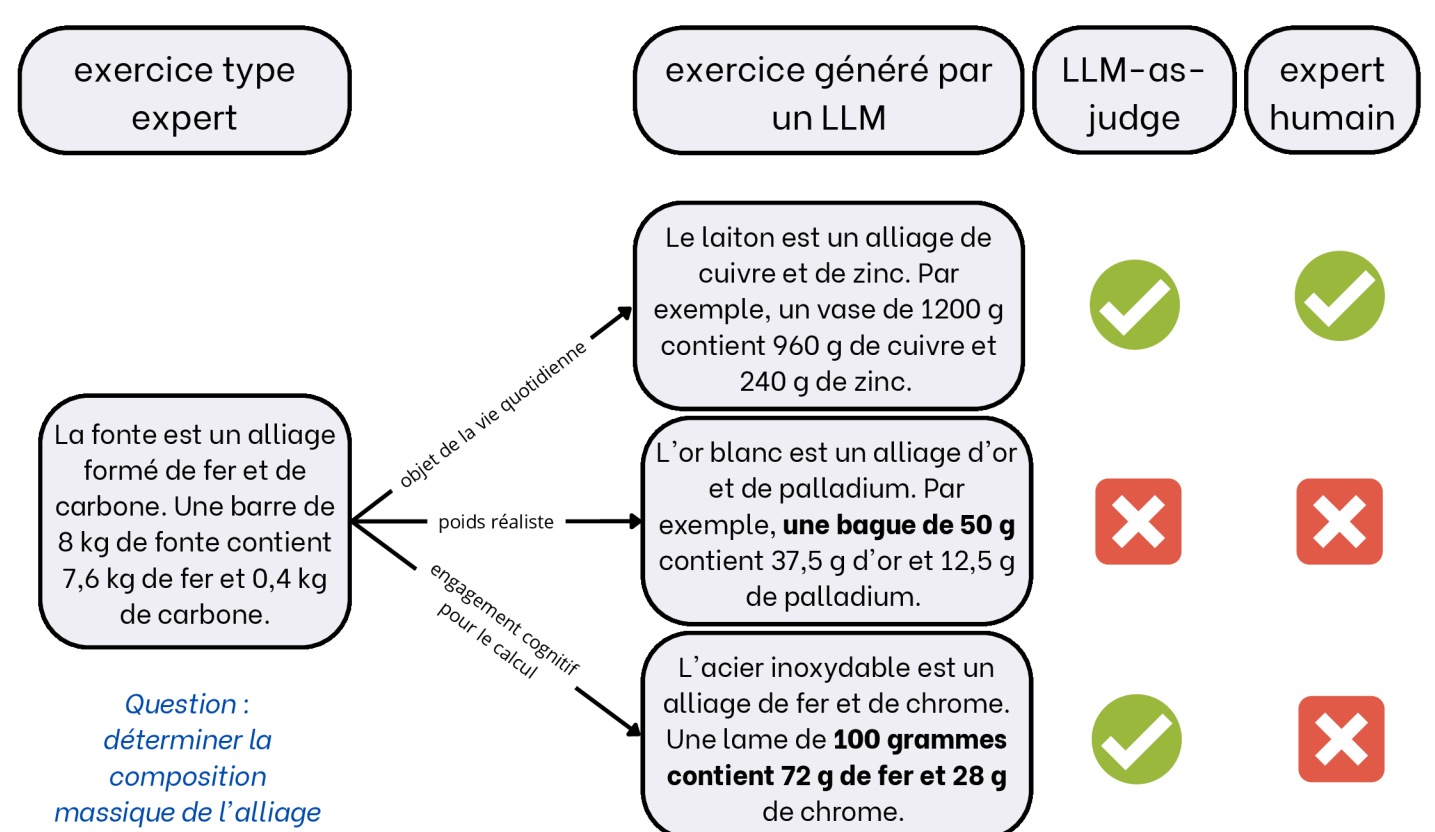
### Contenu:

- Clarté
- Validité pédagogique
- Difficulté adaptée
- Cohérence des éléments de réponses (QCU)
- Vraisemblance du contexte
- Vraisemblance des valeurs

### Feedback:

- Clarté
- Validité pédagogique
- Validité de la réponse
- Complétude

## Exemple d'évaluation humaine vs LLM-as-judge



## Limites & conclusion

- La plupart des LLMs ne sont pas entraînés sur des données éducatives
- Le contenu généré peut être erroné ou ambigu / ne pas correspondre aux objectifs pédagogiques / au niveau attendu
- Usage des LLMs est encadré: ils n'interviennent que pour la génération et non pour la personnalisation du parcours
- Intervention humaine est essentielles à toutes les étapes de la génération du contenu

## Références

CALDERON N., REICHTART R. & DROR R. (2025). The alternative annotator test for LLM-as-a-judge : How to statistically justify replacing human annotators with llms. arXiv : 2501.10970.  
CLÉMENT B., ROY D., OUDEYER P.-Y. & LOPES M. (2015). Multi-armed bandits for intelligent tutoring systems. Journal of Educational Data Mining, 7(2). arXiv : 1310.3174, DOI :10.5281/zenodo.3554667.  
CLÉMENT B., SAUZÉON H., ROY D. & OUDEYER P.-Y. (2025). Improved performances and motivation in intelligent tutoring systems : Combining machine learning and learner choice. arXiv :2402.01669.  
GOTTLIEB J. & OUDEYER P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. Nature Reviews Neuroscience, 19(12), 758–770. DOI : 10.1038/s41583-018-0078-0.  
JURENKA et al. (2024). Towards responsible development of generative ai for education : An evaluation-driven approach. arXiv : 2407.12687.  
ROELL M., DE VULPILLIÈRES C., KNOPS A. & VAGHARCHAKIAN L. (2025a). From intuition to abstraction : Supporting the transition to formal fraction understanding with AI-powered tools. In The Mathematical Cognition and Learning Society (MCLS), Hong Kong. In I. Resnick (Chair), The malleability and utility of informal fraction knowledge from early years contexts through formal schooling.  
ROELL M., DE VULPILLIÈRES C., KNOPS A. & VAGHARCHAKIAN L. (2025b). Leveraging adaptive digital tools to enhance early mathematics learning : Insights from randomized controlled trials. In Conference for Research in Early Childhood Education (CRECE 2025), Hong Kong. In V. Simms (Chair), Unlocking Math Potential : Breakthrough Strategies for All Ages  
TACK A. & PIECH C. (2022). The AI teacher test : Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues. In A. MITROVIC & N. BOSCH, Eds., Proceedings of the 15th International Conference on Educational Data Mining, p. 522–529, Durham, United Kingdom : International Educational Data Mining Society. DOI : 10.5281/zenodo.6853187.